САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА ТЕХНОЛОГИИ ПРОГРАММИРОВАНИЯ

Журавлева Любовь Викторовна

Выпускная квалификационная работа бакалавра

Анализ тональности отзывов пользователей

Направление 010400

Прикладная математика и информатика

Научный руководитель,

старший преподаватель

Попова С. В.

Санкт-Петербург

2016

Содержание

Введение

Анализ тональности текста – это сложный процесс, касающийся выделение полезной субъективной информации из текста. Огромное множество пользовательского контента в интернете появляется с каждым днем. Миллионы пользователей ежедневно высказывают свое мнение о продуктах и услугах в блогах, социальных сетях и других информационных ресурсах. Предоставление надежного извлечения мнения из неструктурированного текста имеет важное значение для коммерческих организаций. С помощью предоставленных данных компании смогут узнать важное для них мнение покупателей, найти невидимые для их глаза недостатки и повысить свой уровень продаж.

Анализ тональности применяется на множестве текстовых документов, содержащих в себе эмоции и оценки определенных объектов, к примеру людей, событий, тем (например, отзывы о фильмах, книгах, продуктах). Анализ тональности предполагает идентификацию сентимента в документе, и в последствии определения его положительной/отрицательной полярности.

Анализ тональности может быть выполнен на различных уровнях – на уровне документа, предложения или аспекта. На уровне документа задача сентимент анализа классифицировать документ, состоящий из множества предложений с точки зрения полярности мнения, выраженного в нем. За основу, часто берется предположение, что в документе выражается мнение о единственном объекте, и в документе не рассматриваются различные точки зрения о данном объекте. [1] На уровне предложения документа задача сентимент анализа классифицировать мнение, и охарактеризовать предложение как положительное, отрицательное или нейтральное. В 2012 году Liu [2] утверждает, что нет никакого различия между уровнем документа и уровнем предложения, и предлагает рассматривать предложения как короткие тексты. Анализ тональности уровня аспекта документа задача сентимент анализа извлекать мнение, выраженные относительно

определённых аспектов объекта. Например, предложение «Качество камеры этого телефона отвратительное, но работает он без подзарядки долго», выражает отрицательное мнение к камере продукта, но выражает положительное мнение относительно работы его аккумулятора.

Как правило, анализ тональности применяется на корпусах текстов, содержащих отзывы. Однако, анализ тональности может быть в том числе применен к новостным статьям [3] или блогам и социальным сетям. Сентимент анализ также применяют чтобы извлечь общественное мнение о различных темах в пределах от фондовых рынков [4] до политических споров [5].

Существуют два основных подхода к задаче автоматического извлечения тональности – подход, основанный на использовании словарей сентиментов и подход, основанный на машинном обучении.

Анализ мнений, основанный на словаре, состоит в анализе тональности слов и фраз представленный в тексте.

Анализ текста в основном ориентируется на использовании прилагательных в качестве сентиментов [7-10] или сочетания прилагательных и наречий [11]. На основе таких слов вместе с их тональностью (положительной или отрицательной) создается словарь сентиментов. Словари сентиметов могут быть созданы вручную [12,13] или автоматически [6,7,14].

Большинство данных методов в исследованиях использовалось для англоязычных словарей. Ряд исследователей из других стран предпринимали попытки создать неанглоязычный словарь различными методами.

Mihalcea [15] предложила два метода для перевода словарей эмоциональных слов (слов сентиментов) на румынский язык. Первый метод состоит в том, что использовались двуязычные словари. Первый – официальный англо-румынский словарь, состоящий из 41,500 слов – будет

использоваться в качестве основного словаря для перевода лексики, второй – взятый с сайта Universal Dictionary, состоящий из 4500 записей, который будет использоваться как дополнение к основному словарю (при отсутствии слов в основном словаре), чтобы перевести английский словарь слов сентиментов, собранный с помощью OpinionFinder [16]. Таким образом был создан словарь, включивший в себя 4983 румынских слова. Второй метод основывается на параллельном корпусе. Корпус на английском языке содержит информацию о типе настроения каждого текста. Далее происходит перевод всех текстов на румынский язык. После попытки тестирования классификатора на текстах, переведенных с английского на румынский язык возникает следующая проблема – после неправильных переводов не распознается ирония, которая определяет тональность текста на английском языке.

Стенберг [17] в 2011 году предлагает свой метод для повышения качества словаря слов сентиментов на испанском языке. Он рассматривает два словаря эмоциональных слов - английский и испанский. Данные словари переводятся на интересующий нас язык с помощью google-переводчика. Новые, не совпадающие записи на испанском языке в дальнейшем добавляются в испанский словарь сентиментов. Исследования проводились для шести языков - итальянского, чешского, арабского, французского, немецкого и русского. Для русского языка объем словаря слов сентиментов составил 966 записей. В открытом доступе данный словарь не найден.

Анализ мнений, основанный на машинном обучении, существенно опирается на результаты из линейной алгебры, мат. анализа, методов оптимизации и теории вероятностей.

Обучение на размеченных данных или обучение с учителем – это наиболее распространенный класс задач машинного обучения. К нему относятся те задачи, где нужно научиться предсказывать некоторую величину для любого объекта, имея конечное число примеров [18], [19]. Это

может быть предсказание уровня пробок на участке дороги, определение возраста пользователя по его действиям в интернете, предсказание цены, по которой будет куплена подержанная машина.

В задаче автоматического извлечения тональности с помощью машинного обучения используются заранее размеченные по тональности коллекции (корпусы) данных, на которых происходит обучение модели, которая в дальнейшем используется для классификации.

Для решения задачи анализа тональности отзывов пользователей в данной квалификационной работе будут использоваться два подхода – основанный на словаре и основанный на машинном обучении. За основу для создания собственного словаря сентиментов будет взят словарь ключевых слов, составленных на основе коллекций отзывов о фотоаппаратах, книгах и фильмах. Полученный словарь сентиментов будет расширен с помощью технологии word2vec. Для демонстрации подхода, основанного на машинном обучении будет использоваться наивный байесовский классификатор. В задаче используется базовый алгоритм машинного обучения, т.к. главная задача состоит в показе возможности улучшить словарь сентиментов с помощью word2vec, а подход основанный на машинном обучении будет использован для демонстрации сравнения с подходом основанном на словарях.

Постановка задачи

Целью данной работы является исследование и разработка метода анализа тональности данных. Для достижения поставленной цели в работе решаются следующие задачи:

1. Исследование и Анализ состояние дел в области и определение основных подходов для решения задачи классификации отзывов по тональности. Изучение основ обработки естественного языка и алгоритмов классификации.
2. Оценка качества классификации по тону на основе предметно-зависимого словаря сентиментов. Построение указанного словаря вручную и исследование возможности его расширения с использованием технологии word2vec.
3. Исследование задачи классификации отзывов по тональности при использовании алгоритмов машинного обучения с учителем.
4. Сравнение и оценка полученных результатов, формирование выводов и рекомендаций по результатам проведённых исследований.

Глава 1. Word2vec

Word2vec - это набор алгоритмов, рассчитывающий векторное представление слов на основе статической информации. Которое реализует два разных подхода для обучения нейронной сети - Continuous Bag of Words (CBOW непрерывный мешок слов) и Skip-gram. Данные архитектуры не используются вместе, а происходит выбор одной из них. При использовании CBOW задача сводится к предсказанию слова на основании слов, находящихся рядом. В случае метода skip-gram при обучении решается обратная задача – на основании одного слова предсказываются контекст (набор близлежащих слов).

В CBOW и Skip-gram используются как подходы в обучении искусственной нейронные сети. Именно с помощью них находятся векторные представления. Изначально каждое слово в словаре обозначается как случайный N-мерный вектор. В процессе обучения классификатора происходит формирование оптимального вектора для каждого слова с помощью двух данных методов.

На рисунке показаны архитектуры методов CBOW и skip-gram. На схеме w(t) – это данное слово, w(t-2), w(t-1) – близлежащие слова.

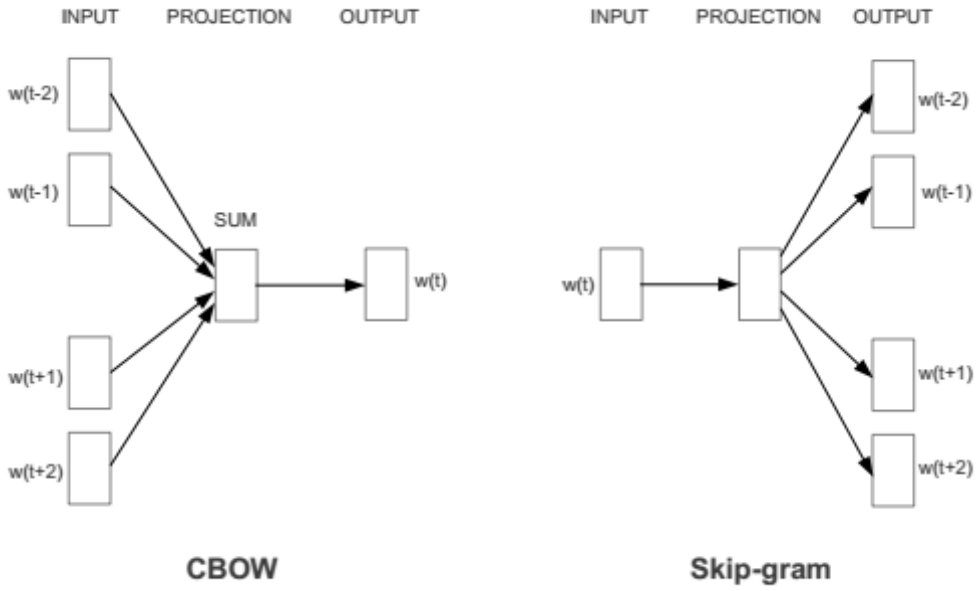

Рисунок 1. Архитектура методов CBOW и skip-gram

В методе CBOW входной слой представляет из себя вектор всех слов, все значения которого равны нулю, кроме одного, по которому мы задаем контекст, этот элемент принимает значение единицы. Этот случай изображен на рисунке 2.

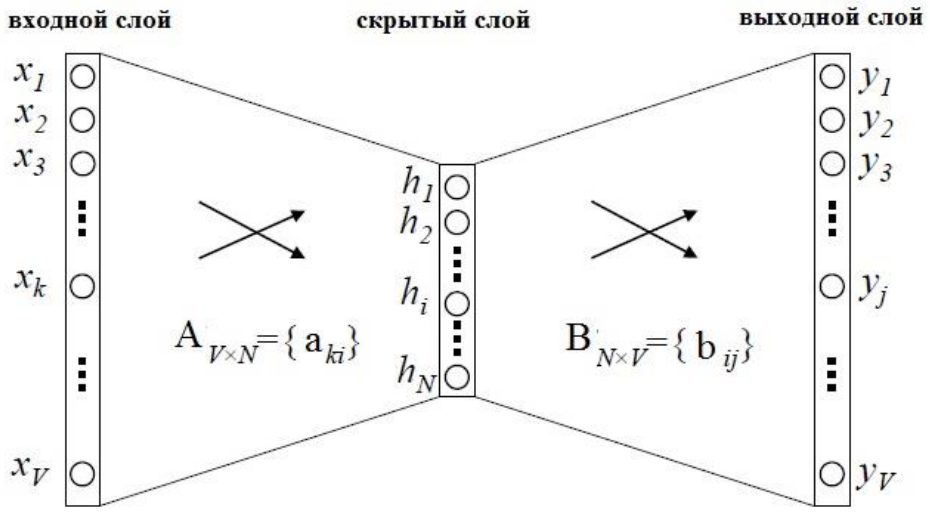

Рисунок 2. Архитектура метода CBOW для случая задания контекста с помощью одного слова.

Если используется несколько слов для задавания контекста, то на вход подается несколько векторов. Данный случай рассмотрен на рисунке 3.

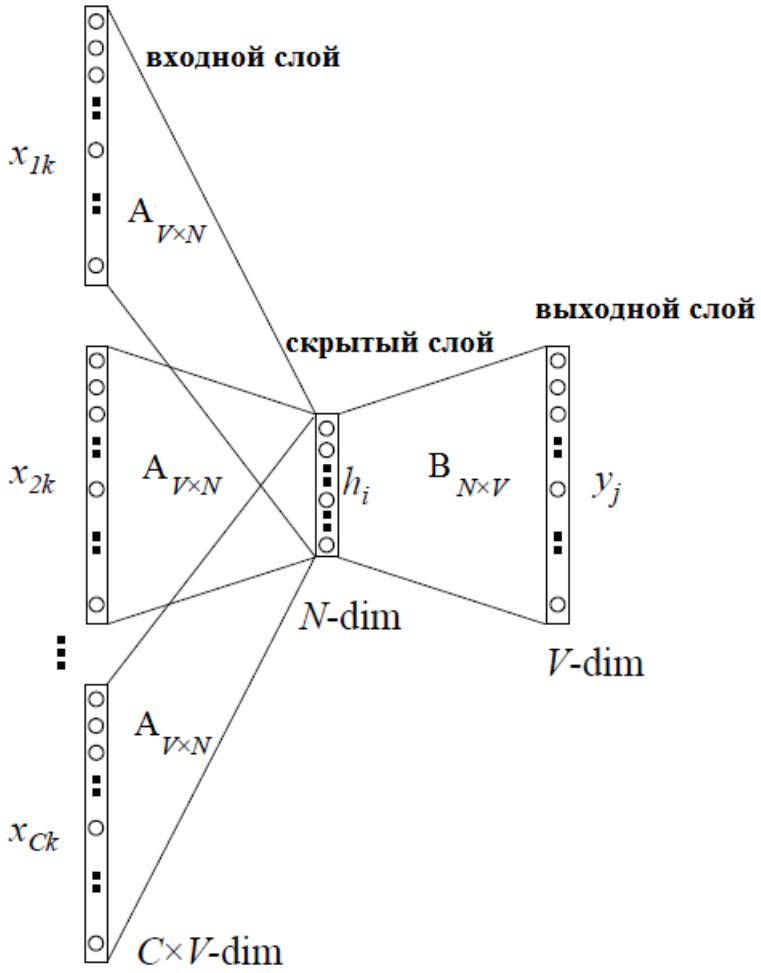

Рисунок 3. Архитектура метода CBOW для случая задания контекста с помощью нескольких слов.

Скрытый слой представляет из себя сумму элементов (векторов) матрицы A размерности V*N скрытых представлений слов - $\sum_{j=1}^{K} A_w(t-j)$.

Слово предсказывается по контексту:

$$p(v|w) = \frac{\exp(B_V^T(\sum_j A_{wj}))}{\sum_{i=1}^{V} \exp(B_i^T(\sum_j A_{wj}))}$$

,где B_v – выходное представление слова (вектор-элемент матрицы B, выходное представление существует для каждого слова).

Входные и выходные представления обучаются в ходе оптимизации целевой функции:

$$F(A, B) = -\sum_{t} \log p(w_t | w_{t-1}, \dots, w_{t-K})$$

В методе skip-gram скрытый входной слой также представляет из себя значение скрытого представления слова. Выходной слой представляет из себя выходное представление слова.Метод skip-gram можно визуально представить, как перевернутую модель CBOW – из слова предсказывается его контекст. Более подробно про данный метод можно прочитать в [20].

Таким образом в word2vec реализуется нейросеть прямого распространения.

Описанный алгоритм, основанный на вычислении относительной частоты слов, задающих соответствующие тематики, хорош тем, что подходит для текстов любой длины. От одного слова и до бесконечности. При этом, как мы знаем, трудно найти достаточно длинный текст одной тематики, часто тематика текста изменяется от его начала к концу. Короткий текст, или сообщение, наоборот, не может охватить множество тем именно за счёт своей краткости. Как итог, получается, что семантический вектор длинного текста отличается признаками нескольких тем от короткого текста, в которых признаков тем намного меньше, но они представлены намного сильнее.

В данной работе использовалась модель национального корпуса русского языка, состоящая из 107 561 399 токена и знающая 281 776 лемм. Данная модель реализовывала структуру Continuous Bag-of-Words с размерностью векторов 300 и размером окна 2.

Глава 2 Словарь сентментов

Вычислительная скорость и эффективность основанных на словаре подходов к анализу мнений делают такие подходы привлекательной альтернативой для извлечения эмоционального контекста из документа.

Не смотря на высокую привлекательность подхода, использование общих предметно не зависимых словарей сентиментов сопряжено с рядом сложностей.

Loughran и McDonald в 2011 году показали наглядную демонстрацию, что применение общего списка слов для определения тональности в сфере финансов приводит к высокому уровню ошибочной классификации. Они обнаружили, что около 0,75 негативных слов в Harvard IV TagNeg словаре, как правило, не негативные в финансовом контексте. К примеру, слова “mine”, “cancer”, “tire” или “capital” часто всего используются для обозначения конкретной отрасли. Эти слова не показывают мнение документа или финансовой новости и их можно отнести к словам, которые не несут смысловую нагрузку. Loughran и McDonald создали собственные списки отрицательных и положительных слов, для сферы учета и финансов.

Данная таблица иллюстрирует различные категории созданного словаря.

Градация	Кол-во слов	Примеры слов
Негативные	2337	termination, discontinued, penalties, misconduct, serious, noncompliance, deterioration, felony
Позитивные	353	achieve, attain, efficient, improve, profitable
Неопределенные	285	approximate, contingency, depend, fluctuate, indefinite, uncertain

Слова с неопределенной тональностью (имеющие двойной смысл)	731	claimant, deposition, interlocutory, testimony
Слабые модальные слова	27	could, depending, might, possibly
Сильные модальные слова	19	always, highest, must, will

Таблица 1 Примеры словаря Loughran и McDonald

Young и Soroka (2011) рассказывают о процессе создания своего словаря для анализа тональности мнений. Их цель была создать словарь на основе объединения уже существующих словарей сентиментов, не ставя под угрозу при этом точность. За основу они взяли словарь - Harvard IV (Stone et al., 1966) и добавили к нему другие положительные и отрицательные слова из Roget's Thesaurus, а также из Colin Martindale's Regressive Imagery. Они удалили нейтральные и неоднозначные слова. Таким образом у них получился словарь, состоящий из 2858 отрицательных записей и 1709 положительных записей.

Вдохновившись этими исследованиями, разработчиками был создан словарь сентиментов WordStat. Этот словарь был фактически разработан, комбинируя отрицательные и положительные слова из словаря Harvard IV, Regressive Imagery dictionary (Martindale, 2003) и Linguistic and Word Count dictionary (Pennebaker, 2007). Была создана утилита, которая использовалась для расширения словаря WordStat. Она автоматически идентифицировала потенциальные синонимы и связные слова, а также любые флективные формы (неправильные формы образования слова). Таким образом они получили 9164 отрицательных и 4847 положительных слов.

Для определения отрицательной тональности они использовали следующие правила:

1) Отрицательные слова, которым не предшествует отрицание (не и др.) в трех словах в том же предложении

2) Положительные слова, которым предшествует отрицание (не и др.) в трех словах в том же предложении

Положительная тональность определялась похожим образом – искалось положительное слово, которому не предшествовало отрицание, и отрицательное слова, которому предшествовало отрицание. Однако по итогам исследования оказалось, что данное последнее правило может даже ухудшить точность измерения тональности, но в некоторых ситуациях данное правило помогало предсказывать положительную тональность.

Рассмотрим еще несколько известных словарей, составленные специально с учетом тональности.

WordNet-Affect

WordNet – это электронный для английского языка, стремящийся дать описание лексики данного языка во всем ее объеме и полноте, разработанный в Принстонском университете [21]. В WordNet базовой единицей взяли не отдельное слово, а синонимический ряд (синсеты), который объединяет слова с похожим значением, и по сути являющимися узлами сети.

С помощью выбора и отнесения к различным эмоциональным понятиям, на основе WordNet был создан WordNet-Affect для английского языка [22]. Были вручную размечены синсеты основных частей речи специальными эмоциональными метками, которые характеризуют различные состояния, выражающие ситуации, которые вызывают эмоции или эмоциональные отклики. Данные метки объединяются, и в дальнейшем разделяются на четыре дополнительных эмоциональных метки: неоднозначная, нейтральная, позитивная и негативная.

Если рассматривать структуру WordNet-Affect, то тезаурус состоит из шести файлов-категорий: печаль, отвращение, удивление, радость, страх, гнев. На данный момент в этом словаре около 2900 синсетов и 4800 слов.

SentiWordNet

SentiWordNet – это словарь, полученный посредством автоматического аннотирования синсетов из WordNet в соответствии с его степенью позитивности, негативности и объективности [23]. Каждая из этих степеней оценивается значением из интервала (0;1), причем все три в сумме должны давать 1.

Процесс создания SentiWordNet состоял из двух шагов:

1. Использовались методы машинного обучения с частичным привлечением учителя. Вначале выбиралось небольшое кол-во синсетов, которые заранее были отмечены вручную. Далее на них обучались несколько классификаторов, которые определяли численные оценки каждого из синсетов. Т.е. в итоге через полученные модели оставшиеся синсеты были размечены.
2. К данным применялась модель случайного блуждания, чтобы установить окончательные оценки объективной, позитивной или негативной составляющей каждого синсета.

Глава 3 Корпус данных

3.1 Отбор данных

Правильный отбор данных для тестирования очень важен. Без тестовых данных невозможно четко и точно определить работоспособность построенного алгоритма.

При определении данных для тестов следует учитывать следующие параметры:

- Объем данных – количество тестовых данных
- Полнота – степень вариации тестовых данных
- Охват – применимость данных для тестирования

Объем

Объем – это количество данных для тестирования. Объем данных – это важный параметр, если данных слишком мало, то они не смогут отразить все аспекты, в моем случае при малом количестве данных невозможно будет построить полноценный словарь, который бы исправно работал в различных ситуациях.

Полнота

Полнота данных означает степень изменения данных для тестирования. Ее можно увеличивать, создавая больше записей.

Охват

Охват – это применимость тестовых данных для проверки верной работы алгоритма. Она связана с объемом и полнотой данных. Большое число данных не означает, что среди них обязательно есть нужные данных. К примеру 15000 отзывов с положительными оценками совершенно не помогут выделить и определить отрицательные отзывы.

3.2 Данные о корпусе

В качестве тестовых данных я воспользовалась коллекцией отзывов о фильмах, которая была взята с портала imhoment.ru. Коллекция была предоставлена НП РОМИП.[24]

Коллекция была представлена в виде xml-кода, отрывок которого приведен ниже:

<row rowNumber="0">

<value columnNumber="0">10</value>

<value columnNumber="1">3</value>

<value columnNumber="2">196076</value>

<value columnNumber="3">23499</value>

<value columnNumber="4">Замечательный фильм, очень рекомендую.

</value>

</row>

Данный отрывок содержит оценку и текст мнения. Корпус данных состоит из 15718 таких отрывков. После предварительной обработки данных было определено, что данный корпус состоит из 1502030 слов.

Было принято решение, считать положительными отзывами те тексты, которые имеют оценку выше пяти баллов, а отрицательными отзывами те тексты, которые имеют оценку ниже 5 баллов.

В таблице 2 приведены некоторые полученные характеристики данных.

Тип мнения	количество мнений	количество полученных слов	средняя длина мнения
Положительный	8108	452620	55.82387
Отрицательный	1928	129258	67.04253

Таблица 2: Характеристики данных

Таким образом получилось 8108 положительных отзывов и 1928 отрицательных отзывов. Т.е. теперь положительные отзывы состояли из 452620

слов, а отрицательные отзывы состояли из 129258 слов. Средняя длина положительного отзыва составила 55.82387 слов, а средняя длина отрицательного отзыва составила 67.04253 слов.

Глава 4 Экспериментальная часть

4.1 Предобработка данных

Так как исходные данные являются необработанным текстом, была проведена предварительная обработка для приведения документов к нормализованному виду.

В своей программе для первоначальной обработки текста я использовала библиотеку Beautiful Soup.

Beautiful Soup - это парсер для синтаксического разбора файлов HTML/XML, написанный на языке программирования Python, который может преобразовать даже неправильную разметку в дерево синтаксического разбора. Он поддерживает простые и естественные способы навигации, поиска и модификации дерева синтаксического разбора.

Для работы конструктору Beautiful Soup требуется документ XML или HTML в виде строки (или открытого файлоподобного объекта). Он произведет синтаксический разбор и создаст в памяти структуры данных, соответствующие документу.

Если обработать с помощью Beautiful Soup хорошо оформленный документ, то разобранная структура будет выглядеть также, как и исходный документ. Но если его разметка будет содержать ошибки, то Beautiful Soup использует эвристические методы для построения наиболее подходящей структуры данных, но работает данная система только для HTML документов, поскольку в XML документах нет фиксированного порядка тегов.

В моем случае я работала с XML документом, в котором содержались отзывы пользователей о фильмах и HTML документом, который был получен при парсинге сайта [25] для расширения словаря сентиментов с помощью word2vec. В первом случае задача состояла в том, чтобы выделить из корпуса

данных информацию для двух массивов – массива оценок и массива непосредственных отзывов.

Так как полученные тексты представляли из себя набор из нескольких предложений, было принято решение избавится от лишних знаков и выделить отдельные слова – для этого текст разбивался на слова регулярным выражением «^[\s\W]*|[^\w]|\s(?=[\W\s]|$)(?u)». Все полученные слова были переведены в нижний регистр.

4.2 Создание словаря

4.2.1 Создания словаря коллекции

Для создания словаря все полученные слова после предобработки были проверены на уникальность (т.е. были проверены на то, встречаются ли они уже в словаре (получаемом с помощью обработки текстов) или нет) и те, которые удовлетворяли данному запросу (отсутствовали в словаре сентиментов) были добавлены в словарь. Также чтобы избавится от слов не несущих информационную пользу я воспользовалась библиотекой NLTK.

Библиотека NLTK содержит в себе список стоп-слов для русского языка. Найденные в этом списке слова, встречающиеся в массиве слов коллекции были удалены из массива. Стоп-слова делают текст тяжелее, слабее, длиннее и не несут смысловую нагрузку, поэтому было принято решение избавиться от них. В задачах классификации было показано, что часто удаление таких слов позволяет повысить качество решаемой задачи[26].

Стоп-слова делятся на несколько категорий: наречия, модальные глаголы, союзы, предлоги, междометия, частицы и другие.

Таким образом из массива слов из положительно оцененных текстов ушло 121260 словоупотреблений стоп-слов, из массива слов из отрицательно оцененных текстов ушло 36112 словоупотреблений стоп-слов.

Далее был создан на основе полученного массива словарь, который насчитывал 69496 слов из положительных отзывов и 26660 слов из отрицательных отзывов.

4.2.2 Создания словаря сентиментов

Второй словарь был получен с помощью ручной обработки.

За основу был взят словарь Четверкина [27], который состоял из ключевых слов, взятых из корпусов о фильмах, книгах и фотоаппаратах и составлял 5000 слов без разделения по тональности.

Для каждого слова была проведена оценка вручную на тональность (положительную и отрицательную), слова, которые было невозможно отнести к той или иной тональности исключались. Таким образом специально для мета-области фильмов было получено 1926 слов с разделением по тональности. Из них 1002 было с положительной тональностью и 924 с отрицательной тональностью.

На основе word2vec (с помощью модели «Ruscorpora and Russian Wikipedia», представленной на сайте [25]), который позволяет находить похожие слова, которые семантически являются близкими, был расширен словарь сентиментов.

К примеру, для слова прекрасный семантически близкими окажутся следующие слова:

слово
замечательный
великолепный
чудесный
восхитительный
превосходный
шикарный
удивительный

Таблица 3. Семантически близкие слова к слову «прекрасный»

Таким образом, словарь сентиментов расширился до 2392 слов с положительной тональностью и 1586 слов с отрицательной тональностью.

4.3 Используемая система для экспериментов

Все вычисления производились на ноутбуке Acer Aspire V 15 Nitro. Использовался процессор 2,20 ГГц Intel Core i5-5200U, 6 Гб оперативной памяти. Операционная система Windows 8.1.

4.4 Использованные методы определения тональности

4.4.1 Методы, основанные на машинном обучении

Наивный байесовский классификатор

Байесовский подход к классификации основан на теореме, утверждающей, что если плотности распределения каждого из классов известны, то искомый алгоритм можно выписать явно в аналитическом виде. На практике плотности распределения, как правило не известны. Их приходится восстанавливать, основываясь на обучающей выборке.

Наивный байесовский классификатор – это специальный частный вариант байесовского классификатора. Основан на применении теоремы Байеса со строгими (наивными) предположениями о независимости (изменение одной величины не влияет на изменение другой величины).

В основе NBC (Naïve Bayes Classifier) находится теорема Байеса.

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

где,

P(c|d) – вероятность того, что документ d принадлежит классу c

P(d|c) – вероятность встретить документ d среди всех документов класса c

P(c) – безусловная вероятность встретить документ класса c в корпусе документов

P(d) – безусловная вероятность документа d в корпусе документов

Цель классификации – понять к какому классу относится объект, в нашем случае понять положительный или отрицательный у нас отзыв.

Байесовский классификатор использует оценку апостериорного максимума (грубо говоря класса с максимальной вероятностью)

$$c_{map} = \arg\max_{c \in C} \frac{P(d|c)P(c)}{P(d)}$$

Таким образом необходимо рассчитать вероятность для «положительного» и «отрицательного» класса и выбрать тот класс, который обладает максимальной вероятностью.

Так как P(d) является константой, и никаким образом не изменяется, то при расчете его можно опустить. Получим такую формулу

$$c_{map} = \arg\max_{c \in C}[P(d|c)P(c)] \ (1)$$

В естественном языке вероятность появления слова сильно зависит от контекста. Байесовский классификатор представляет документ как набор слов вероятности, которые условно не зависит друг от друга. Данный подход еще иногда называется bag of words (мешок слов). Исходя из этого предположения условная вероятность документа аппроксимируется произведением условных вероятностей всех слов, входящих в документ.

$$P(d|c) \approx P(w_1|c)P(w_2|c)\ldots P(w_n|c) = \prod_{i=1}^{n} P(w_i|c)$$

Подставив полученное выражение в (1) получаем:

$$c_{map} = \text{argmax}_{c \in C}[P(c) \prod_{i=1}^{n} P(w_i|c)]$$

Проблема арифметического переполнения

При достаточно большой длине документа придется перемножать большое количество очень маленьких чисел. Для того чтобы при этом избежать арифметического переполнения снизу зачастую пользуются свойством логарифма произведения logab=loga+logb. Так как логарифм функция монотонная, ее применение к обоим частям выражения изменит только его численное значение, но не параметры при которых достигается максимум. При этом, логарифм от числа близкого к нулю будет числом отрицательным, но в абсолютном значении существенно большим чем исходное число, что делает логарифмические значения вероятностей более удобными для анализа. Поэтому, переписываем нашу формулу с использованием логарифма.

$$c_{map} = \text{argmax}_{c \in C}[\log P(c) + \sum_{i=1}^{n} \log P(w_i|c)] \quad (2)$$

Основание логарифма может быть любым. Можно использовать также натуральный логафирм. В своей работе я использовала логарифм с основанием 10.

Оценка параметров Байесовской модели

Оценка вероятностей P(c) и $P(w_i|c)$ происходит на обучающем множестве. Вероятность класса мы можем оценить следующим образом:

$$P(c) = \frac{D_c}{D}$$

где, D_c – кол-во документов, принадлежащих рассматриваемому классу c, а D – общее количество документов в обучающей выборке.

Оценка может осуществляться несколькими способами. Рассмотрим способ multinomial bayes model.

$$P(w_i|c) = \frac{W_{ic}}{\sum_{i' \in V} W_{i'c}} \quad (3)$$

где, W_{ic} – сколько кол-во раз i-ое слово встречается в документах класса c, V – список всех уникальных слов (словарь рассматриваемого корпуса)

Проблема неизвестных слов

Если на этапе классификации вам встретится слово которого вы не видели на этапе обучения, то значения W_{ic} , а следственно и $P(w_i|c)$ будут равны нулю. Это приведет к тому что документ с этим словом нельзя будет классифицировать, так как он будет иметь нулевую вероятность по всем классам. Избавиться от этой проблемы путем анализа большего количества документов не получится. Вы никогда не сможете составить обучающую выборку, содержащую все возможные слова включая неологизмы, опечатки, синонимы и т.д. Решением данной проблемы неизвестных слов является аддитивное сглаживание (сглаживание Лапласа).

Идея состоит в том, чтобы увеличить частоты слов, для этого прибавляется единица к частоте каждого слова.

$$P(w_i|c) = \frac{W_{ic} + 1}{\sum_{i' \in V}(W_{i'c} + 1)}$$

Таким образом слова, которых не было в обучающем множестве получат маленькую, но не нулевую вероятность.

Окончательная формула (2) будет выглядеть следующим образом.

$$c_{map} = \text{argmax}_{c \in C}[\log \frac{D_c}{D} + \sum_{i=1}^{n} \log \frac{W_{ic}+1}{|V|+\sum_{i' \in V} W_{i'c}}] \quad (4)$$

Для реализации классификатора перепишем (4) в более удобный вид:

$$\log \frac{D_c}{D} + \sum_{i \in Q} \log \frac{W_{ic} + 1}{|V| + L_c}$$

где,

Dc— количество документов в обучающей выборке принадлежащих классу c;

D — общее количество документов в обучающей выборке;
|V| — количество уникальных слов во всех документах обучающей выборки;
Lc— суммарное количество слов в документах класса с в обучающей выборке;
Wic — сколько раз i-ое слово встречалось в документах класса с в обучающей выборке;
Q – множество слов классифицируемого документа (включая повторы).

4.4.2 Методы, основанные на словарном подходе

Алгоритм для словарного подхода

В качестве определения тональности для словарного подхода, было принято использовать метод, описанный Peter D. Turney в 2002 году [28]. Идея заключается в следующем: каждое слово текста рассматривается на наличие в словаре, как слово, несущее положительную или отрицательную тональность. Если слово встречается в словаре, как слово с положительным весом, то счетчик для слов, несущих положительную тональность увеличивается. Аналогично счетчик для слов несущих отрицательную тональность, увеличивается, если слово имеет отрицательный вес.

Тональность текста определяется большим количеством того или иного счетчика.

4.5 Методы оценки качества алгоритма

Для того чтобы понять, насколько хорошо построенный алгоритм работает с данными, необходима численная метрика его качества.

Точность и полнота

Точность (precision) – это доля документов, которые действительно принадлежат классу относительно всех документов, который классификатор отнес к этому классу.

Полнота(recall) – это доля найденных системой документов, принадлежащих классу относительно всех документов рассматриваемого класса в тестовой выборке.

Данные значения можно рассчитать на основе таблицы контингентности

Категория i		Экспертная оценка	
		положительная	отрицательная
оценка системы	положительная	TP	FP
	отрицательная	FN	TN

Таблица 4. Таблица контингентности.

Где

TP – истинно-положительное решение

FP – ложно-положительное решение

FN – ложно-отрицательное решение

TN – истинно-отрицательное решение

Тогда полнота и точность вычисляются следующим образом

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Высокая точность и полнота означают качественное построение модели.

F-мера

F-score (F-мера) это метрика, которая объединяет информацию о точности и полноте алгоритма. F-score представляет собой гармоническое среднее между точностью и полнотой. Данная метрика стремится к нулю, если полнота и точность стремятся к нулю

$$\text{F} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.6 Результаты

Для точного определения работы подхода, основанного на словарях было принято решение разделить исходный корпус данных на 30 выборок, состоящих из 500 отзывов, собранных случайным образом.

Используя метод, описанный в параграфе 4.4.2 для словаря сентиментов, полученного ручной обработкой по тональности из словаря ключевых слов, были получены следующие результаты:

выборка	1	2	3	4	5
precision	0.92857	0.89912	0.79629	0.81655	0.79661
recall	0.55667	0.54959	0.64759	0.62707	0.59119
f-score	0.69606	0.68219	0.71428	0.70937	0.67870

Таблица 5. Таблица результатов словаря сентиментов для выборок 1-5

выборка	6	7	8	9	10
precision	0.82278	0.92760	0.85833	0.83721	0.81081
recall	0.57863	0.53525	0.55675	0.51136	0.50991
f-score	0.67944	0.67881	0.67541	0.63492	0.62608

Таблица 6. Таблица результатов словаря сентиментов для выборок 6-10

выборка	11	12	13	14	15
precision	0.92704	0.88679	0.88889	0.88207	0.81696
recall	0.54135	0.62500	0.54026	0.51800	0.54790
f-score	0.68354	0.73323	0.67205	0.65270	0.65591

Таблица 7. Таблица результатов словаря сентиментов для выборок 11-15

выборка	16	17	18	19	20
precision	0.85771	0.88281	0.86695	0.92523	0.91964
recall	0.59452	0.59317	0.54742	0.49748	0.54787
f-score	0.70226	0.70957	0.67110	0.64705	0.68667

Таблица 8. Таблица результатов словаря сентиментов для выборок 16-20

выборка	21	22	23	24	25
precision	0.92373	0.91379	0.92647	0.89868	0.90232
recall	0.58602	0.55208	0.52646	0.55738	0.49743
f-score	0.71710	0.68831	0.67140	0.68803	0.64132

Таблица 9. Таблица результатов словаря сентиментов для выборок 21-25

выборка	26	27	28	29	30
precision	0.89316	0.88537	0.93776	0.91775	0.88235
recall	0.56639	0.60705	0.59162	0.52736	0.55556
f-score	0.69320	0.72026	0.72552	0.66983	0.68182

Таблица 10. Таблица результатов словаря сентиментов для выборок 26-30

Справедливо предположение о том, что данные выборки имеют нормальное распределение, проверенное с помощью критерия согласия Пирсона [29]. Таким образом, представив данные в виде диаграммы можно заметить высокую точность. Недостаточно высокую полноту можно объяснить малым словарным запасом словаря, который состоит из 2000 слов.

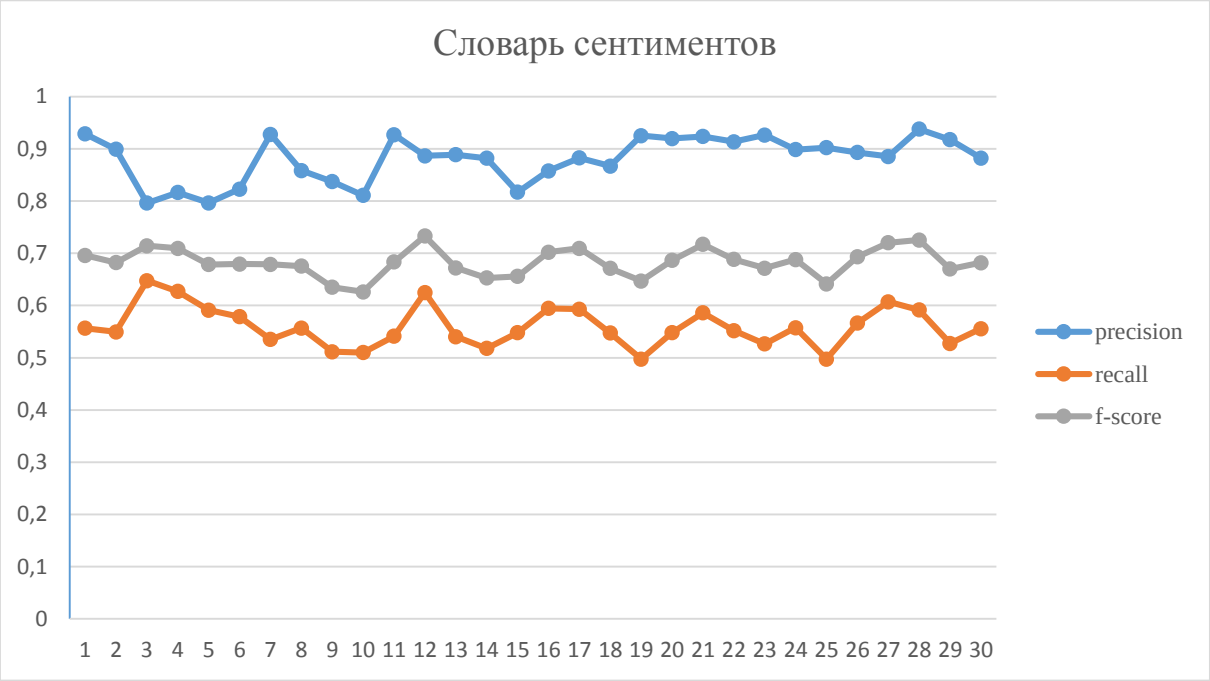

Диаграмма 1. Результаты словаря сентиментов

Рассмотрим, теперь отдельно оценки определения позитивных и негативных отзывов. Результаты получились следующими:

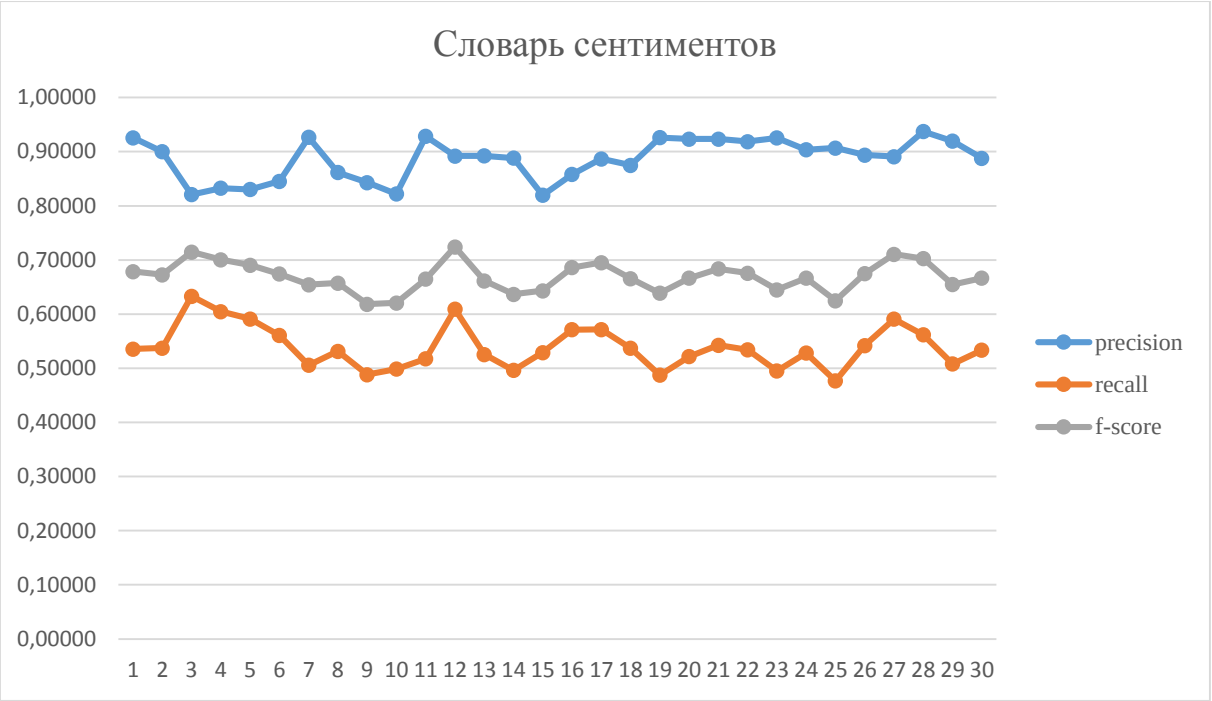

Диаграмма 2. Результаты оценки определения позитивных отзывов для словаря сентиментов.

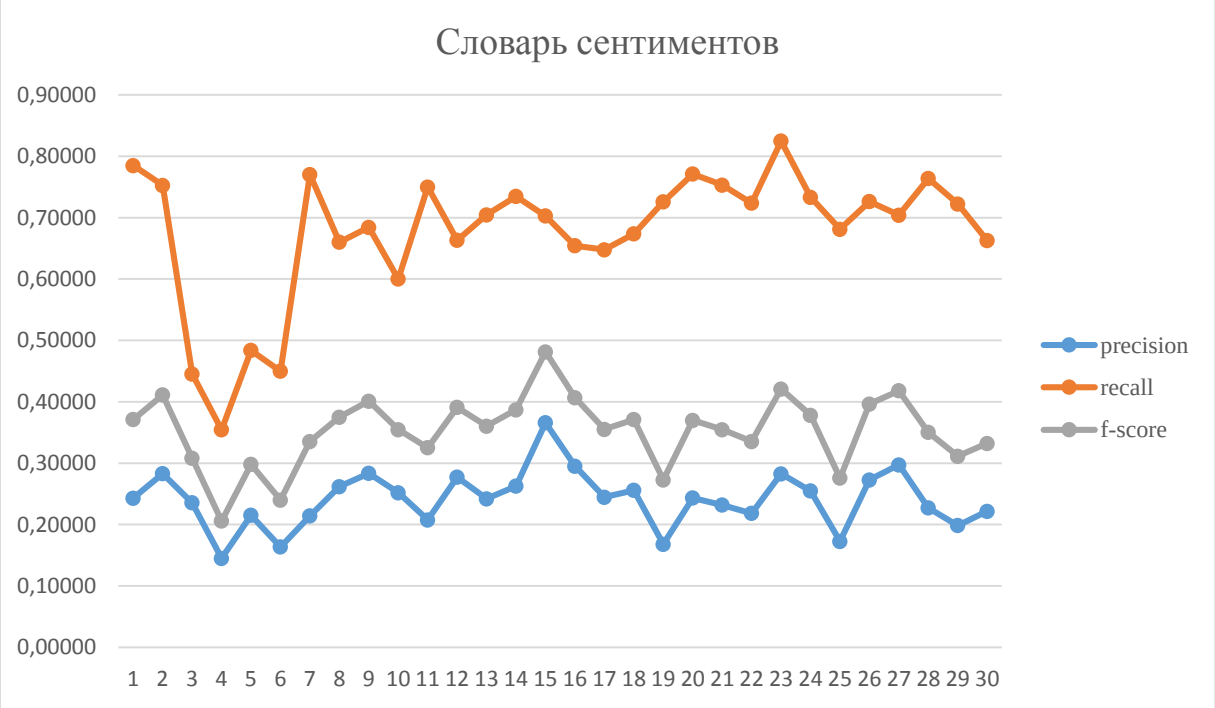

Диаграмма 3. Результаты оценки определения отрицательных отзывов для словаря сентиментов.

Результаты оценки словаря сентиментов показали более высокую точность определения положительных отзывов, чем отрицательных, и более высокий общий результат.

Используя метод, описанный в параграфе 4.4.2 для расширенного с помощью word2vec словаря сентиментов, были получены следующие результаты:

выборка	1	2	3	4	5
precision	0.89520	0.86750	0.767726	0.81840	0.76595
recall	0.75314	0.73726	0.945783	0.93370	0.90566
f-score	0.81805	0.79710	0.84750	0.87225	0.82997

Таблица 11. Таблица результатов словаря сентиментов для выборок 1-5

выборка	6	7	8	9	10
precision	0.81521	0.87539	0.82335	0.79822	0.79765
recall	0.89020	0.73368	0.74324	0.76420	0.77053
f-score	0.85106	0.79829	0.78125	0.78084	0.78386

Таблица 12. Таблица результатов словаря сентиментов для выборок 6-10

выборка	11	12	13	14	15
precision	0.90797	0.83776	0.85965	0.83132	0.78154
recall	0.74185	0.83776	0.76364	0.76454	0.76048
f-score	0.81655	0.83776	0.80880	0.79654	0.77086

Таблица 13. Таблица результатов словаря сентиментов для выборок 11-15

выборка	16	17	18	19	20
precision	0.82386	0.83714	0.85	0.89966	0.87009
recall	0.79452	0.76903	0.73713	0.67588	0.76596
f-score	0.80893	0.80164	0.78955	0.77188	0.81471

Таблица 14. Таблица результатов словаря сентиментов для выборок 16-20

выборка	21	22	23	24	25
precision	0.89130	0.87692	0.87205	0.86218	0.89809
recall	0.77150	0.74219	0.72145	0.73497	0.72308
f-score	0.82709	0.80395	0.78963	0.79351	0.80114

Таблица 15. Таблица результатов словаря сентиментов для выборок 21-25

выборка	26	27	28	29	30
precision	0.85449	0.86163	0.9	0.87425	0.87353
recall	0.74797	0.74255	0.77749	0.72637	0.78571
f-score	0.79769	0.79767	0.83427	0.79348	0.82729

Таблица 16. Таблица результатов словаря сентиментов для выборок 26-30

Данные в виде диаграммы получаются следующими:

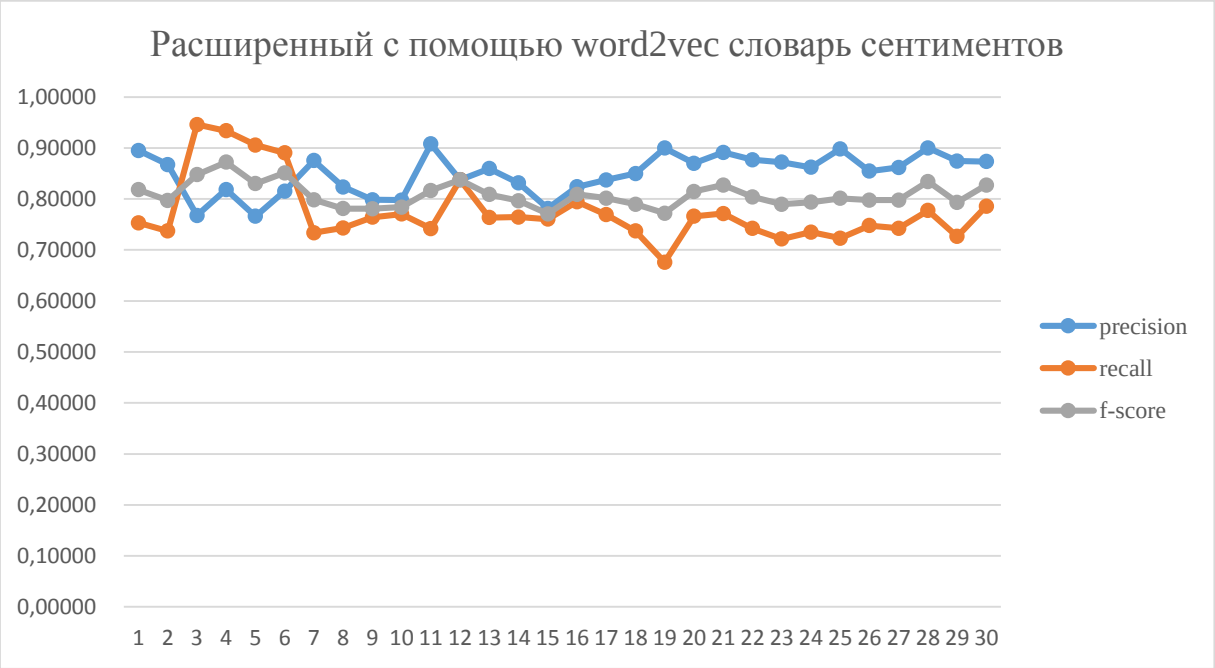

Диаграмма 4. Результаты расширенного с помощью word2vec словаря сентиментов.

Значения f-score были проверены на нормальное распределение. Справедливо предположение о том, что данные выборки имеют нормальное распределение.

Если провести сравнение точностей словаря сентиментов размеченного вручную, и этого же словаря сентиментов расширенного с помощью word2vec можно заметить достаточно малое уменьшение точности на тысячную долю процентов. Это обуславливается тем, что некоторые слова, полученные с помощью технологии word2vec могут не нести ту же тональность, что и слово, с помощью которого искались эти слова.

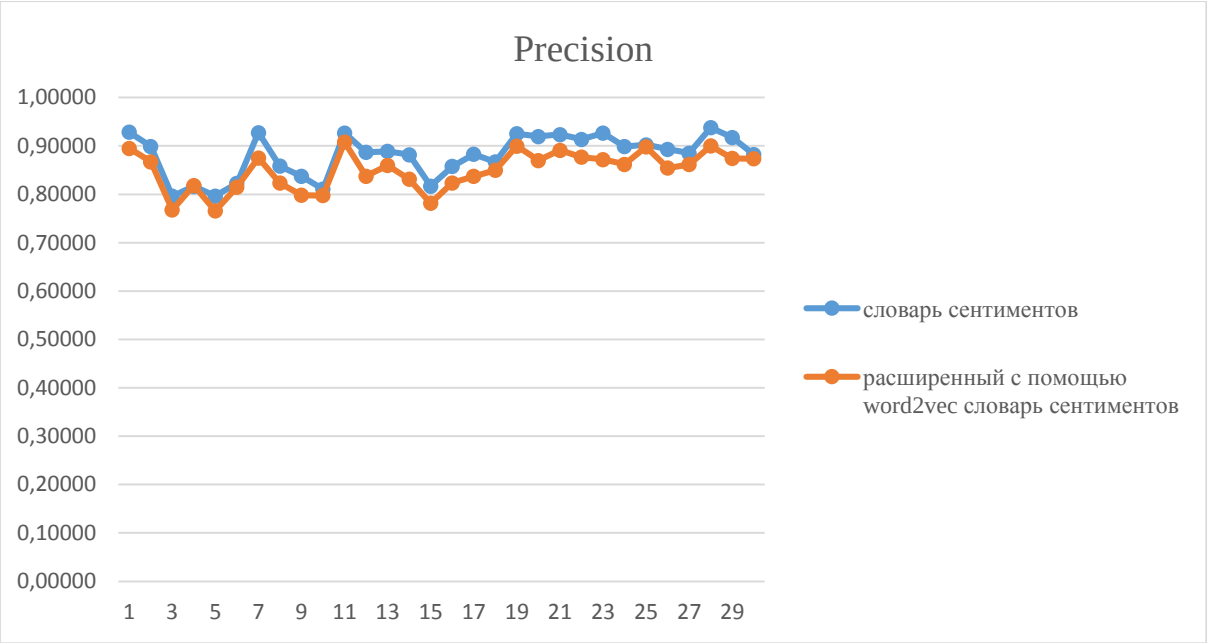

Диаграмма 5. Результаты сравнения точности

Т.к. словарь был значительно расширен, т.е. количество слов возросло практически в два раза, сложно не заметить очень большое увеличение полноты.

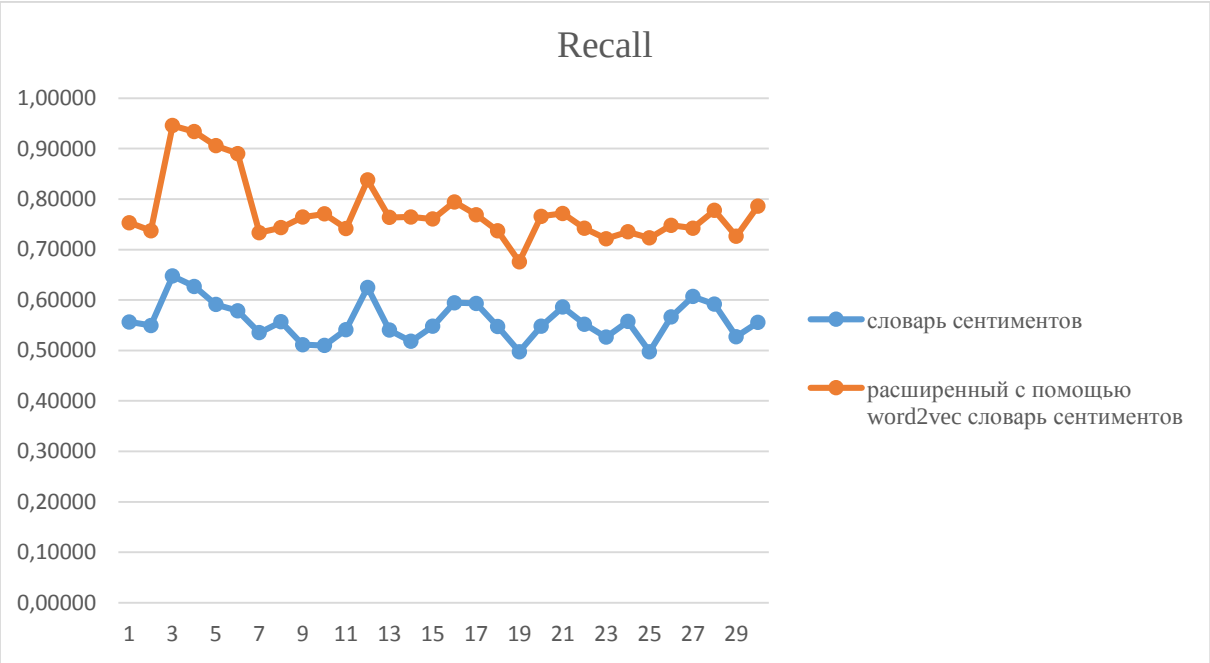

Диаграмма 6. Результаты сравнения полноты

Метрика, объединяющая информацию о точности и полноте алгоритма, также возросла.

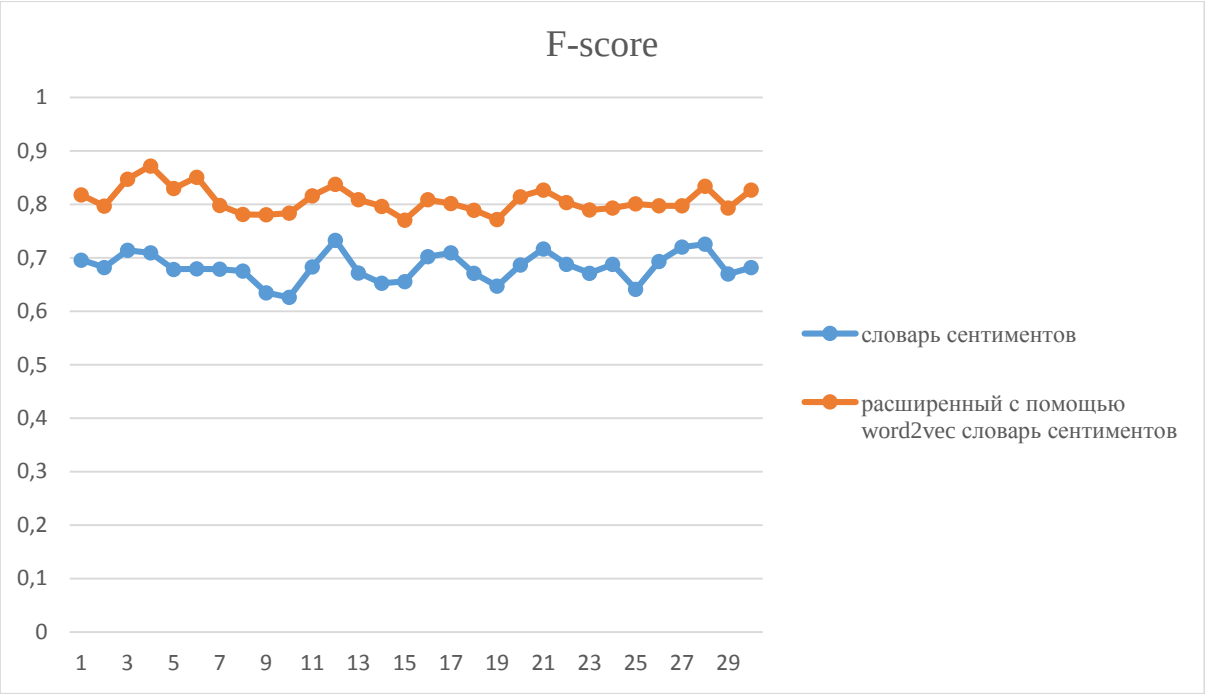

Диаграмма 7. Результаты сравнения f-меры

Рассмотрим, теперь отдельно оценки определения позитивных и негативных отзывов. Результаты получились следующими:

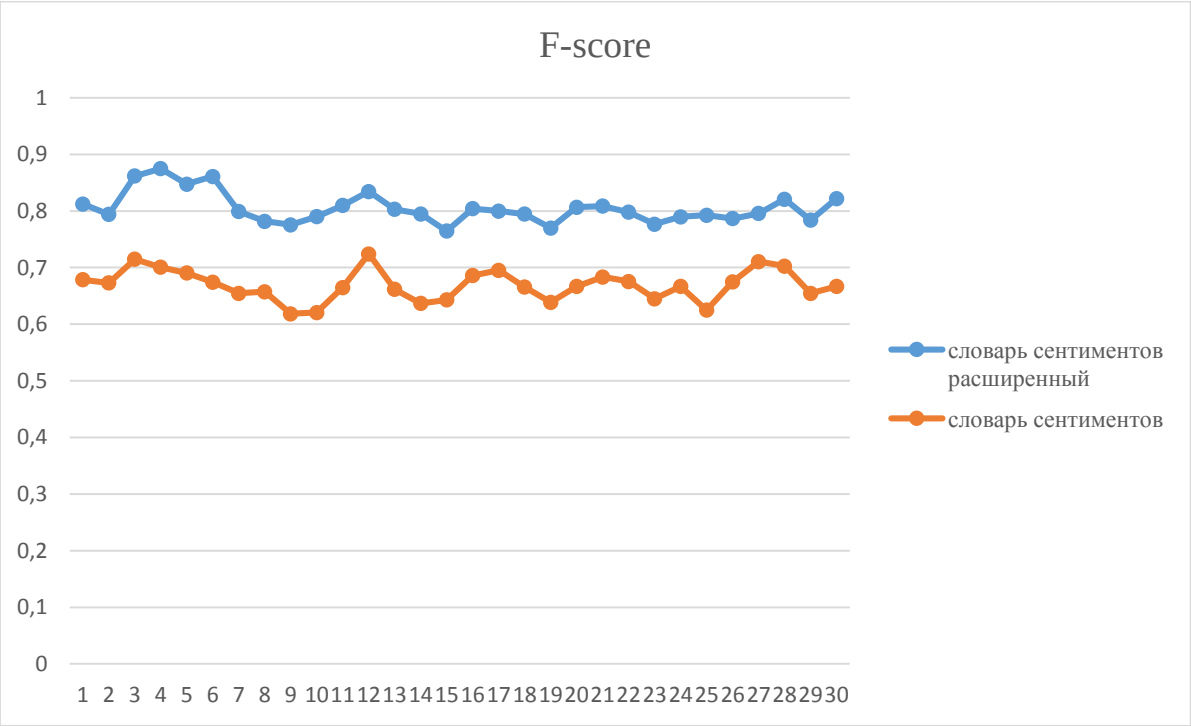

Диаграмма 8. Результаты f-меры расчета положительных отзывов для словаря сентиментов.

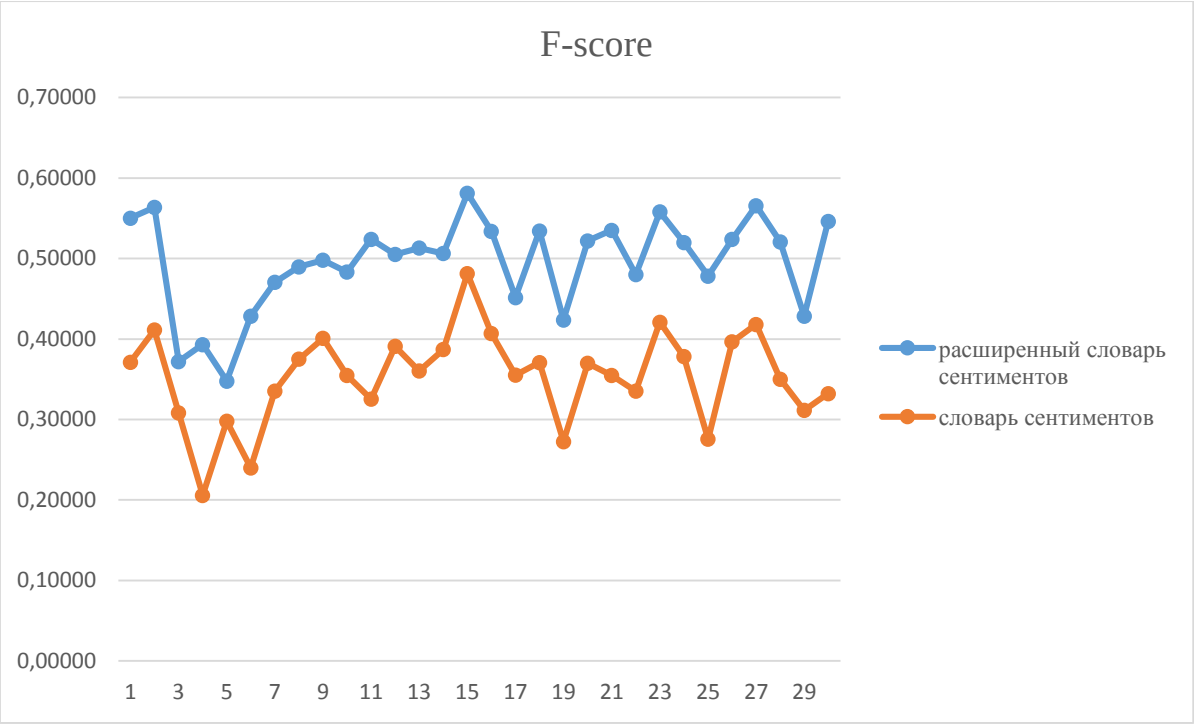

Диаграмма 9. Результаты f-меры расчета отрицательных отзывов для словаря сентиментов.

Таким образом словарный подход в данной квалификационной работе показывает достаточно высокую точность. Использование технологии word2vec имеет положительную тенденцию, увеличивая значительно полноту и общий результат работы представленной модели. Качество расчета положительных отзывов все также выше, чем у отрицательных.

Для точного определения работы подхода, основанного на машинном обучении было принято решение разделить исходный корпус данных на две части – обучающую коллекцию, которая составила 7500 отзывов и тестовую коллекцию, которая была разделена на 30 выборок по 250 отзывов, собранных случайным образом. Данные были представлены в виде xml-кода.

Результаты получились следующими:

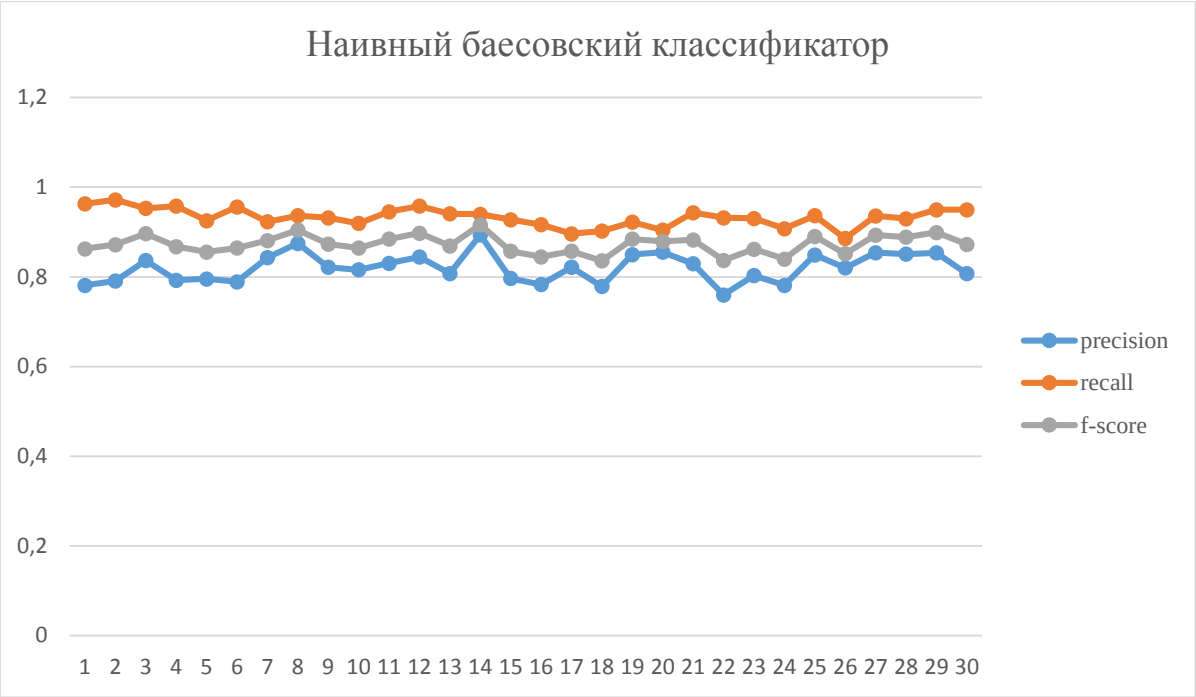

Диаграмма 10. Результаты наивного байесовского классификатора

Справедливо предположение о том, что данные выборки имеют нормальное распределение. Очень высокая полнота объясняется тем, что словарь, полученный при обработке 7500 отзывов содержит в себе достаточное большое количество слов, для определения тональности текста.

Для произведения сравнения двух подходов – основанного на машинном обучении (использовался словарь коллекции) и основанного на словаре (использовался словарь сентиментов, расширенный с помощью word2vec) было создано 30 выборок по 250 отзывов из 7500 отзывов, которые не использовались при обучении наивного байесовского классификатора. Результаты получились следующими:

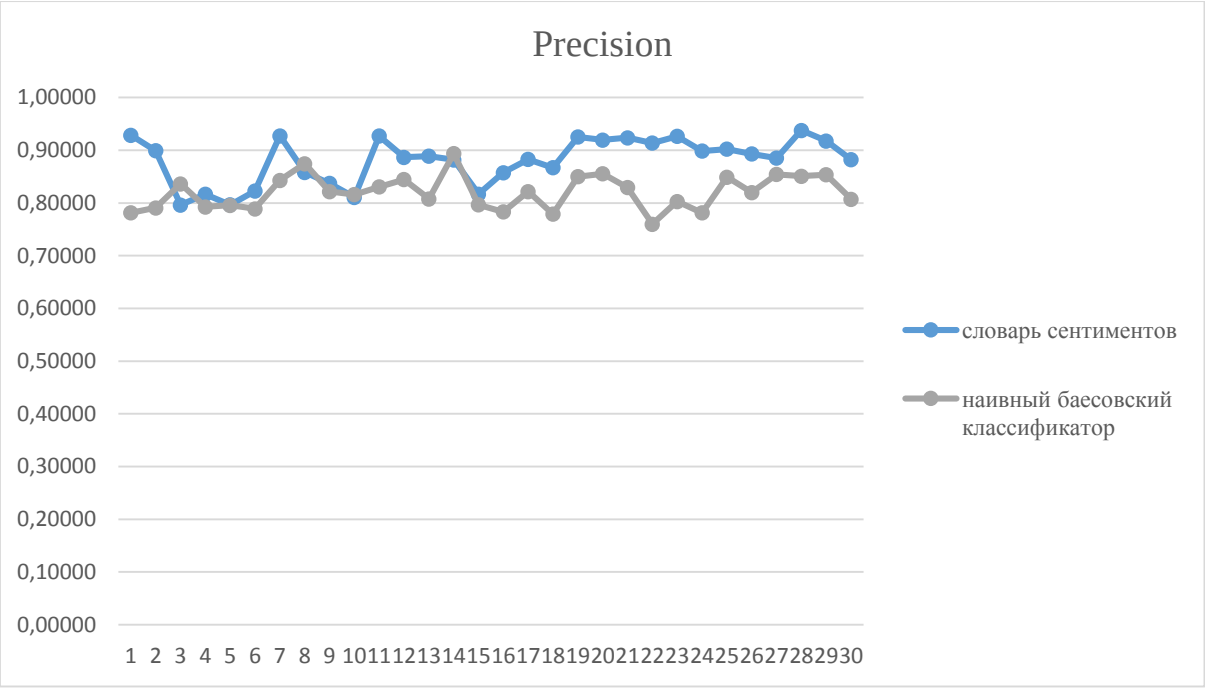

Диаграмма 11. Результаты сравнения точности

Не смотря на небольшой объем словаря сентиментов, данный подход не только не уступает в точности подходу, основанному на машинном обучении, но и превосходит в большинстве случаев.

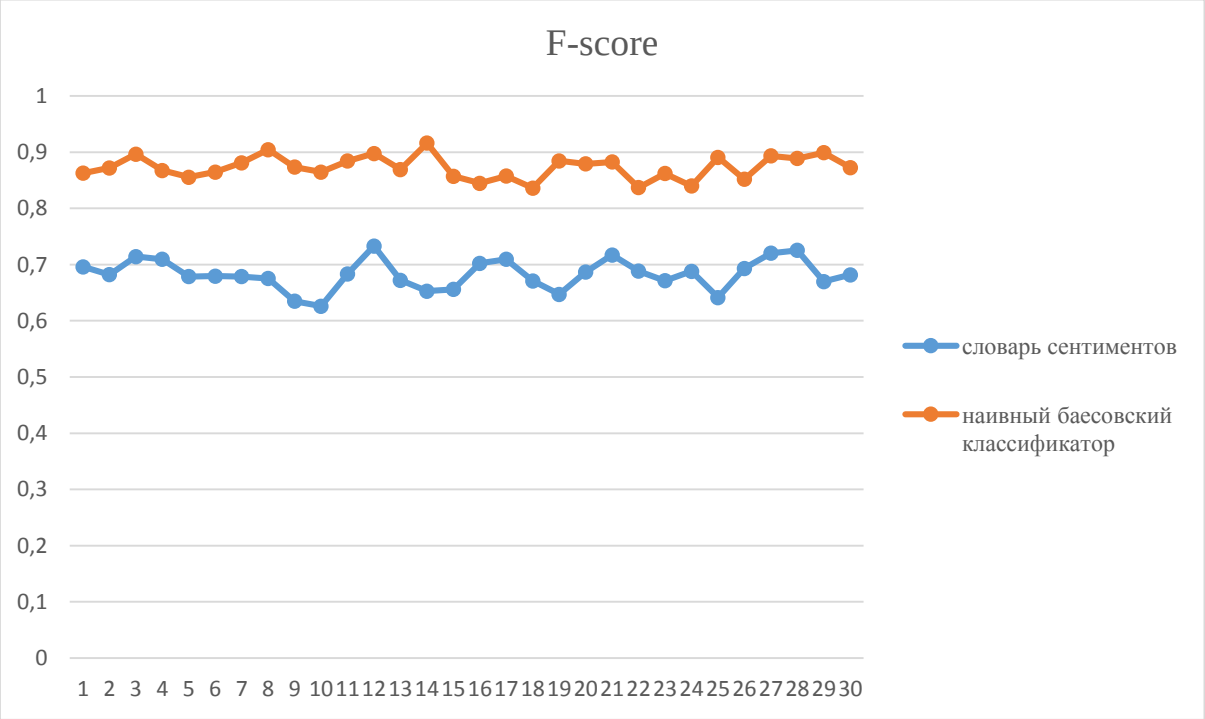

Диаграмма 12. Результаты сравнения f-score

Так как словарь находится на начальной стадии разработки, он еще недостаточно велик и не рассматривает всевозможные слова и исходы, чтобы гарантировать для себя достаточно высокую полноту, поэтому дает наименьший итоговый результат по сравнению с подходом, основанном на машинном обучении.

Заключение

Целью данной работы являлось исследование и разработка метода анализа тональности данных. Для достижения поставленных целей были решены несколько задач. В рамках данной выпускной квалификационной работы была рассмотрена проблема определения тональности отзывов о фильмах. Было проведено исследование состояния дел в области и были определены подходы для решения задачи: с использованием словарей сентиментов и с использованием машинного обучения. Были разобраны разнообразные подходы, которые использовались для создания словаря сентиментов для других языков.

Было представлено исследование задач классификации отзывов по тональности с использованием двух подходов – основанного на словаре и основанного на машинном обучении. Был вручную построен словарь сентиментов, на основе которого были проведены исследования возможности его расширения с помощью технологии word2vec, которые привели к положительному результату – произошло повышения полноты и общего результата определения тональности. Было произведено сравнение оценки качества работы двух подходов на 30 выборках, созданных на основе 7500 отзывов, не входящих в обучающую выборку для наивного байесовского классификатора. Исследование показало более высокую точность подхода, основанного на словаре. При расширении словаря с помощью word2vec было отмечено высокое повышение полноты словаря.

Были определены перспективы развития - в дальнейшем было бы интересно расширить свой собственный словарь сентиментов на основе правил и рассмотреть вариант расширения словаря сентиментов, переведя «Harvard IV» и построить его расширенную версию. Также хотелось бы провести исследование используя как единицу словаря не одно слово, а синонимический ряд и полученные результаты представить в магистерской работе.

Список литературы

1. B. Pang, L. Lee, and S. Vaithyanathan, in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02 (Association for Computational Linguistics, 2002) p. 79-86.
2. B. Liu, Synthesis Lectures on Human Language Technologies 5, 1 (2012).

3. T. Xu, Q. Peng, and Y. Cheng, Knowledge-Based Systems 35, 279 (2012).

4. M. Hagenau, M. Liebmann, and D. Neumann, Decision Support Systems 55, 685 (2013).

5. I. Maks and P. Vossen, Decision Support Systems 53, 680 (2012).

6. P. D. Turney, in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,ACL '02 (Association for Computational Linguistics, 2002) pp. 417-424.
7. V. Hatzivassiloglou and K. R. McKeown, in Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics (Association for Computational Linguistics, 1997) p. 174-181.
8. J. Wiebe, in Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth
Conference on Innovative Applications of Artificial Intelligence (AAAI Press, 2000) pp. 735-740.

9. M. Hu and B. Liu, in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04 (ACM, 2004) p. 168-177.
10. M. Taboada, C. Anthony, and K. Voll, in Proceedings of the Fifth International Conference on Language
Resources and Evaluation, LREC '06 (2006) p. 427-432.

11. F. Benamara, C. Cesarano, A. Picariello, D. R. Recupero, and V. S. Subrahmanian, in Proceedings of
International Conference on Weblogs and Social Media, ICWSM '10 (2007).

12. M. Taboada, J. Brooke, M. Tofloski, K. Voll, and M. Stede, Computational linguistics 37, 267 (2011)

13. R. M. Tong, in Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification,
Vol. 1 (2001) p. 6.

14. P. D. Turney and M. L. Littman, ACM Transactions on Information Systems 21, 315 (2003).

15. Mihalcea R., Banea C. and Wiebe J. (2007). Learning multilingual subjective language via cross-lingual projections. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 976–983, Prague, Czech Republic.

16. Wiebe J. and Riloff E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In Proceedings of CICLing 2005. pp. 486-497.

17. Steinberger J., Lenkova P., Ebrahim M., Ehrmann M., Hurriyetogly A., Kabadjov M., Steinberger R., Tanev H., Zavarella V. and Vazquez S. (2011). Creating Sentiment Dictionaries via Triangulation. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pp. 28–36.

18. R. Prabowo and M. Thelwall, Journal of Informetrics 3, 143 (2009).
19. Q. Ye, Z. Zhang, and R. Law, Expert Systems with Applications 36, 6527 (2009).
20.Xin Rong, "word2vec Parameter Learning Explained." arXiv preprint arXiv:1411.2738, 2014
21. Fellbaum C. "WordNet: An Electronic Lexical Database" MIT Press, 1998.

22. Strapparava C., Valitutti A. WordNet-Affect: An affective extension of WordNet Proceedings of LREC, 2004, pp. 1083–1086.
23. Baccianella S. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of LREC, 2010, pp. 2200–2204.
24. http://romip.ru/
25. http://ling.go.mail.ru/dsm/en/similar
26. Калимондаев М.Н., Пак А.А., Нарынов С.С. . Нейросетевой метод семантического вероятностного вывода в задаче улучшения релевантности результатов поискового запроса. Проблемы информатики 2014,стр 82-86
27. http://www.cir.ru/SentiLexicon/ProductSentiRus.txt
28. Turney, P.D. (2002), Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, Pennsylvania, 417-424.
29. http://termist.com/bibliot/stud/stepnov/081_2.htm